

UNITED STATES DISTRICT COURT  
SOUTHERN DISTRICT OF NEW YORK

Rio Tinto plc,

Plaintiff,

-against-

Vale S.A., Benjamin Steinmetz, BSG  
Resources Limited, VBG–Vale BSGR  
Limited aka BSG Resources (Guinea) Ltd.  
aka BSG Resources Guinée Ltd, BSG  
Resources Guinée SARL aka BSG  
Resources (Guinea) SARL aka VBG-Vale  
BSGR, Frederic Cilins, Mamadie Touré,  
and Mahmoud Thiam,

Defendants.

USDC SDNY  
DOCUMENT  
ELECTRONICALLY FILED  
DOC#  
DATE FILED: 3/2/15

14 Civ. 3042 (RMB) (AJP)

**STIPULATION AND ORDER RE:  
USE OF PREDICTIVE CODING IN DISCOVERY**

WHEREAS, the Plaintiff Rio Tinto and the Defendant Vale (collectively the “Parties” and each a “Party”) in the above captioned litigation (“Action”) agree to the use of predictive coding for the search, review, and production of documents in this Action and to enter a stipulation (“Stipulation”) to memorialize their agreement;

IT IS HEREBY STIPULATED AND AGREED, by and between the undersigned, as attorneys of record for the Parties, as follows:

1. Definitions

- (a) “Precision” means the fraction of documents identified as likely responsive by the Predictive Coding Process that are in fact responsive.<sup>1</sup>

---

<sup>1</sup> Definitions of “Precision” and “Recall” are adapted from the “Grossman-Cormack Glossary of Technology-Assisted Review,” 7 Fed. Cts. L. Rev. 1, 25, 27 (2013).

- (b) “Predictive Coding Process” means the use by the Parties of Predictive Coding Software to categorize documents into those that are likely responsive and those that are likely non-responsive.
- (c) “Predictive Coding Software” means the software a Party elects to use to perform the Predictive Coding Process.
- (d) “Recall” means the fraction of responsive documents that are identified as likely responsive by the Predictive Coding Process.
- (e) “Statistically Valid Sample” means a random sample of sufficient size and composition to permit statistical extrapolation with a margin of error of +/- 2% at the 95% confidence level.<sup>2</sup>

2. Scope of this Stipulation

(a) The procedures described in this Stipulation govern the use of predictive coding to assist in the production of documents by the Parties to this Action. In this Stipulation, predictive coding shall mean and refer to a process for selecting and ranking a collection of documents using a computerized system that incorporates the decisions that lawyers have made on a smaller set of documents and then applying those decisions to the remaining universe of documents.

(b) Nothing in this Stipulation shall prevent the Party responding to discovery (the “Responding Party”) from using other search, review, or coding methodologies in

---

<sup>2</sup> The size of the sample will vary depending upon several factors, and shall be calculated using the formula

$$n = \frac{X^2 \cdot N \cdot P \cdot (1-P)}{(ME^2 \cdot (N-1)) + (X^2 \cdot P \cdot (1-P))}$$
, where, ME is the margin of error; X is the Confidence Level (1.96 for a 95% Confidence Level); P is judgment of richness, N is the population and n is sample size. Where richness is not reasonably estimable, 0.5 may be used. Based on a Confidence Level of 95%, richness of 0.5, a Population of 1,000,000, and a margin of error of 2%, the resulting sample size is 2,395 documents.

addition to, or in place of, predictive coding to help identify documents that are responsive to the document requests from the Party seeking discovery (the “Requesting Party”).

3. Initial Disclosure of Information about Predictive Coding

(a) Prior to the commencement of any review using predictive coding, the Responding Party shall disclose to the Requesting Party in writing its intention to use predictive coding and the following information:

- (i) The name, publisher, version number, and a description of the Predictive Coding Software and Predictive Coding Process;
- (ii) The name and qualifications of the person who will oversee the implementation of the predictive coding process (the “Technical Expert”);
- (iii) A description of the documents to be subjected to predictive coding (the “Document Universe”), including:
  - (1) Custodian / Source<sup>3</sup>;
  - (2) Data types (e.g., email, electronic documents, etc.);
  - (3) The number of documents in the Document Universe, in total and for each Custodian / Source;
- (iv) The responsiveness categories into which the Document Universe is to be categorized (the “Responsiveness Categories”).

(b) The Parties shall meet and confer to address any questions or disputes about the selection of the Predictive Coding Software, the Technical Expert, the Document Universe, and the Responsiveness Categories, and the Responding Party shall make its

---

<sup>3</sup> These terms have the definitions set forth in the ESI Protocol (Dkt. No. 82).

Technical Expert reasonably available to address questions about the technical operation of the Predictive Coding Software.

4. Predictive Coding Methodology

(a) Culling the Document Universe. If the Responding Party determines it to be reasonable and appropriate, the Responding Party may use search terms and other criteria (the “Culling Criteria”) to reduce the volume of the Document Universe. If it does so, the Responding Party shall promptly:

(i) Disclose in writing to the Requesting Party the Culling Criteria used and the number of documents removed by the Culling Criteria (the “Excluded Documents”);

(ii) Review a Statistically Valid Sample from the Excluded Documents, disclose the size of that sample set, and produce any responsive, non-privileged documents the Responding Party identifies;

(iii) Meet and confer with the Requesting Party, if requested, to address any questions or disputes about the reasonableness and appropriateness of the Culling Criteria.

(b) Control Set Review. To aid in the Predictive Coding Process, and to determine the prevalence of responsive information from the Document Universe, the Responding Party shall review a Statistically Valid Sample of documents from the Document Universe (the “Control Set”). Prior to the commencement of Seed Set Identification, see 4(c) below, the Responding Party shall disclose the results of the review of the Control Set to the Requesting Party, including the number of documents in the Control Set and the number of documents that were coded for each of the Responsiveness Categories during the review of the Control Set. The Responding Party shall produce all

non-privileged documents reviewed in the Control Set, and for each document disclose the Responsiveness Categories, if any, to which it is responsive. The Requesting Party shall raise any disputes regarding how the documents were coded for each of the Responsiveness Categories in the Control Set within five (5) business days of the production of 4,000 documents or fewer or ten (10) business days of the production of more than 4,000 documents. The parties agree to meet and confer in good faith over any such disputes. All non-responsive documents produced from the Control Set shall be deemed "Highly Confidential" under the terms of the Stipulated Protective Order (Dkt. No. 81), shall be used only for the purpose of evaluating the accuracy of the document coding, and shall be promptly returned or destroyed after review by the Requesting Party and the resolution of any disputes.

(c) Seed Set Identification. The Responding Party may use any reasonable method, including, but not limited to, search terms, to identify a set of documents to be used to initially train the Predictive Coding Software (the "Seed Set"). Prior to commencement of Training, see 4(d) below, the Responding Party shall disclose to the Requesting Party in writing a description of the size of the Seed Set and the methodology used to identify it. The Responding Party shall produce all non-privileged documents and disclose for each document the Responsiveness Categories, if any, to which it is responsive. The Requesting Party shall raise any disputes regarding how the documents were coded within five (5) business days of the production of 4,000 documents or fewer or ten (10) business days of the production of more than 4,000 documents. The parties agree to meet and confer in good faith over any such disputes. All non-responsive documents produced from the Seed Set shall be deemed "Highly Confidential" under the terms of the Stipulated Protective Order (Dkt. No. 81), shall be used only for the purpose of evaluating



the accuracy of the document coding, and shall be promptly returned or destroyed after review by the Requesting Party and the resolution of any disputes.

(d) Training Sets. The Responding Party may use any reasonable method to train the Predictive Coding Software. Upon completion of the training, the Responding Party shall disclose in writing the results of the training to the Requesting Party, including, to the extent reasonably available, the number of documents reviewed, the number of documents coded for each of the Responsiveness Categories during training, the number of documents identified as likely responsive by the Predictive Coding Process, and the estimated rates of Recall and Precision with their associated error margins. The Responding Party shall produce all non-privileged documents used to train the Predictive Coding Software and disclose for each document the Responsiveness Categories, if any, to which it is responsive. The Requesting Party shall raise any disputes regarding how the documents were coded within ten (10) business days of their production, and the parties agree to meet and confer in good faith over any such disputes. All non-responsive documents produced shall be deemed "Highly Confidential" under the terms of the Stipulated Protective Order (Dkt. No. 81), shall be used only for the purpose of evaluating the accuracy of the document coding, and shall be promptly returned or destroyed after review by the Requesting Party and the resolution of any disputes.

(e) Uncategorized Documents. The Responding Party shall disclose the number of documents the Predictive Coding Software is unable to evaluate for any reason, including the unavailability of machine-readable text or documents that could not be ranked, and review such documents in their entirety in order to identify responsive, non-privileged documents for production.

(f) Validation Set. Prior to production, the Responding Party shall review a Statistically Valid Sample of documents in the Document Universe that are categorized as likely non-responsive by the Predictive Coding Process (the “Purported Non-Responsive Documents”) in order to determine the prevalence of responsive documents that are contained therein. Prior to production, the Responding Party shall disclose to the Requesting Party in writing the number of Purported Non-Responsive Documents, the size of the Validation Set, the number of documents identified as responsive during the review of the Validation Set, and the implied rate of Recall. The Responding Party shall produce any responsive, non-privileged documents it identifies during the review of the Validation Set.

5. General Provisions.

(a) The Parties hereby agree to meet and confer in good faith over any disputes that might arise with respect to the terms and conditions of this Stipulation or any other aspects relating to discovery. The Responding Party agrees to make its Technical Expert reasonably available to the Requesting Party for questions about its use of predictive coding. Should the Parties be unable to resolve their disputes on any issues stemming from the use of predictive coding set forth in this Stipulation, they shall submit those issues to the Court for resolution.

(b) Notwithstanding the provisions set forth in this Stipulation, the Parties respectively reserve their rights regarding the instant discovery process. This includes, but is not limited to, the Requesting Party’s right to object to the efforts of the Responding Party to search for, review, and produce information in response to the Requesting Party’s document requests; and the Responding Party’s right to withhold information pursuant to

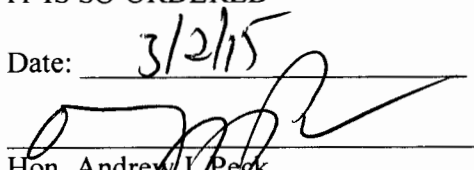
the objections it previously interposed in response to the Requesting Party's document requests.

**BY ECF**

*copy ECF. Del Land  
for Ben*

IT IS SO ORDERED

Date: 3/2/15

  
Hon. Andrew J. Peck  
United States Magistrate Judge

HON. ANDREW J. PECK  
United States Magistrate Judge  
Southern District of New York

---

William A. Burck  
Eric C. Lyttle  
Michael J. Lyle  
Stephen M. Hauss  
QUINN EMANUEL URQUHART & SULLIVAN, LLP  
777 6th Street NW, 11th floor  
Washington, DC 20001  
[williamburck@quinnemanuel.com](mailto:williamburck@quinnemanuel.com)  
[ericlyttle@quinnemanuel.com](mailto:ericlyttle@quinnemanuel.com)  
[mikelyle@quinnemanuel.com](mailto:mikelyle@quinnemanuel.com)  
[stephenhauss@quinnemanuel.com](mailto:stephenhauss@quinnemanuel.com)

*Counsel for Plaintiff Rio Tinto plc.*

---

Jonathan I. Blackman  
Lewis J. Liman  
Boaz S. Morag  
CLEARY GOTTlieb STEEN & HAMILTON LLP  
One Liberty Plaza  
New York, NY 10006  
[jblackman@cgsh.com](mailto:jblackman@cgsh.com)  
[lliman@cgsh.com](mailto:lliman@cgsh.com)  
[bmorag@cgsh.com](mailto:bmorag@cgsh.com)

*Counsel for Defendant Vale S.A.*